Comment

Supplementary information to:

Don't ignore genetic data from minority populations

A Comment published in *Nature* **585**, 184–186 (2020)
https://doi.org/10.1038/d41586-020-02547-3

Chief Ben-Eghan, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V. Grant, Charles Dupras & Simon Gravel

This Supplementary information comprises:

1. Supplementary methods, analysis, results and references (this document)
2. Supplementary data tables 1–4 (separate spreadsheet)

Supplementary information
Don't ignore genetic data from minority populations

Chief Ben-Eghan[1,3,*], Rosie Sun[2,3,*], Jose Sergio Hleap[4], Alex Diaz-Papkovich[3,5], Hans Markus Munter[1,3], Audrey V. Grant[1,3,6], Charles Dupras[7], and Simon Gravel[1,3,†]

[1]Department of Human Genetics, McGill University, Montreal, QC
H3A 0G1, Canada.

[2]Department of Pharmacology and Therapeutics, McGill University, Montreal, QC H3G 1Y6, Canada.

[3]McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada.

[4]SHARCNET, University of Guelph, Guelph, ON
N1G 2W, Canada.

[5]Quantitative Life Sciences Program, McGill University, Montreal, QC, H3A 0G1 Canada.

[6]Alan Edwards Center for Research on Pain, McGill University, Montreal, QC, H3A 0G1 Canada.

[7]Centre of Genomics and Policy, McGill University and Genome Quebec Innovation Centre, Montreal, QC, H3A 0G1, Canada.

[*]contributed equally

[†]email:simon.gravel@mcgill.ca

1. Population labels in Genome Wide Association Studies

We distinguished among *MIN* populations using label choices made by each cohort. In UKB, for example, we considered identity category 'Black and Black British' as a group (*N* = 8,066) and pooled 'Asian and Asian British' and 'Chinese' categories in a single group (*N* = 11,456), despite the heterogeneity within and among groups. No label choice captures the complexity of population structure in ancestry, culture, and environment. However, combining labels in this way allowed us to compare related representation patterns across multiple studies (see Table S1 and Table S2 below for the breakdown of participants from the UKB and HRS).

Table S1: Self-reported ethnic origins of UK Biobank participants and individuals in the UK (all aged 40-69 years)

Self-reported ethnicity [a]	Participants	UK Population 2001	UK Population 2011
White [b]	472,837 (94.6%)	19,085,322 (94.5%)	21,133,317 (91.3%)
Black /Black British [c]	8,066 (1.6%)	302,073 (1.5%)	565,777 (2.4%)
Mixed [d]	2,958 (0.6%)	82,389 (0.4%)	191,085 (0.8%)
Asian/Asian British, and Chinese [e]	11,456 (2.3%)	664,055 (3.3%)	1,107,159 (4.9%)
Other Ethnic group	4,560 (0.9%)	64,468 (0.3%)	149,274 (0.6%)

[a] Excludes 2,778 UK Biobank participants aged 40–69 years with missing data on ethnicity or responded "prefer not to answer" or "do not know."

[b] Includes White British, White Irish, and "other White background".

[c] Includes Caribbean, African, and "other Black background".

[d] Includes White and Black Caribbean, White and Black African, White and Asian, and "other mixed ethnic background".

[e] Includes Indian, Pakistani, Bangladeshi, "Any other Asian background" and Chinese.

Table adapted from Reference 1

Table S2: Self-reported ethnicity of genotyped individuals in the HRS 2016

Ethnicity	Hispanic Status	Participants
Black	Hispanic/ Other	32 (0.25%)
Black	Not Hispanic/ N/A	1,620 (13%)
Other	Hispanic/ Mexican American	90 (0.7%)
Other	Hispanic/ Other	89 (0.7%)
Other	Not Hispanic/ N/A	189 (1.5%)
White	Hispanic/ Mexican American	615 (4.9%)
White	Hispanic/ Other	376 (3.0%)
White	Hispanic/ Type Unknown	1 (0.008%)
White	Not Hispanic/ N/A	9,442 (75.8%)
Total		12,454

2. Literature Review – Methods & Results

We searched the GWAS Catalog [2] with the keyword "UK Biobank" and found 22 articles. We excluded one study which was a gene-environment interaction study. We reviewed 21 articles which conducted a GWAS or meta-analysis using UKB data.

Only 1 out of 21 studies included individuals of all ethnicities in UKB. The other studies restricted their analysis to only *MAJ* individuals in UKB. One study included Asian populations from a different cohort in a primary meta-analysis, but only used *MAJ* populations from UKB in the replication analysis [3]; another study excluded data from *MIN* participants in the UKB for the primary analysis, but included *MIN* populations from other cohorts in the replication [4]. The remaining 18 studies focused exclusively on *MAJ* participants from the UKB. (see Supplementary Data Table 1)

To validate our findings, we queried other databases including Elsevier, Scopus, biomed/Pubmed, biorxiv, and arxiv for articles that mentioned "GWAS", "genome-wide association studies", and "UK Biobank". This generated 1,265 results. Using an online literature review tool, Rayyan QCRI [5], we skimmed through the abstracts for studies that conducted a GWAS based on UKB data until we reached a pre-specified goal of 20 articles, none of which was in the GWAS Catalog list. Only 1 out of 20 studies included data from participants of all ethnicities in UKB, while the rest discarded *MIN* data in the analysis (see Supplementary Data Table 2).

To verify that these findings were not peculiar to UKB only, we used another cohort, HRS and queried the online HRS publication database [6] using the keyword "GWAS" and found 40 results. 17 out of 40 studies from the query results conducted a GWAS or meta-analysis using the HRS data. 11 out of 17 studies included participants of all ethnicities in HRS, while 6 limited their analysis to *MAJ* populations. (see Table S3 below and Supplementary Data Table 3 for details).

Table S3: Proportion of GWAS articles that included *MIN* data in analysis.

	Proportion
UKB (GWAS Catalog)	1/21 (4.7%)
UKB (other)	1/20 (5.0%)
HRS	11/17(64.7%)

3. Proportion of Majority (*MAJ)* and Minority (*MIN*) UK Biobank (UKB) populations Analyzed in the 21 studies queried

In this section, we describe the approach used to estimate the proportion of minority population data used in the GWAS analyses. We used data from the 21 UK Biobank studies we queried from the GWAS catalog.

For each article index $j \in \{1, \ldots, 21\}$, let n_j be the number of individuals from UKB included in the corresponding GWAS analysis, n_j^{MAJ}, the number of individuals from MAJ, and n_j^{MIN} the number of individuals from MIN included in the study. All studies reported n_j^{MAJ}. n_j^{MIN} was taken to be zero if the study stated that if focused on individuals of MAJ, otherwise we used the explicit numbers reported in the study. We then computed the sum of analyzed samples as

$$n_j = n_j^{MAJ} + n_j^{MIN}$$

$$N_{total} = \sum_{j=1}^{21} n_j$$

$$N_{MIN} = \sum_{j=1}^{21} n_j^{MIN}.$$

The effective proportion of MIN individuals at time of analysis is simply $\frac{N_{MIN}}{N_{total}} = 0.0006 = 0.06\%$. This proportion is lower than what we might expect if we simply multiplied the proportion of MIN participants in the UKB (5.4%) by the proportion of studies including the MIN participants (1/21), since the one study that included MIN participants focused on a small subset of the entire UKB.

4. A simple GWAS analysis

To illustrate a minimal stratified analysis, we performed a GWAS of eosinophil cell count in the UKB as part of a broader effort on the genetics of asthma [7]. Eosinophil production is a hallmark of atopic diseases including asthma, and a higher eosinophil blood count is systematically observed across asthma study populations [8].

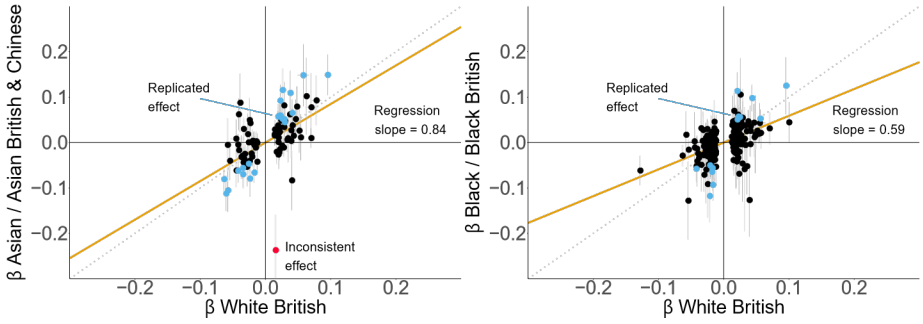
We modelled the eosinophil cell count phenotype by adjusting for the following covariates: age, sex, smoking status, assessment centre where the blood samples were collected, machine counter that processed the blood samples, month, day of the week, time inside the day that the samples were collected, self-reported ethnic background of the individuals, height, weight and 10 principal components. We then normalized the resulting residuals by applying rank-based inverse normal transformation for each population.

Starting with the *MAJ* population (here individuals who self-identify as 'White British' only), we ran standard GWAS quality control pipeline (QC) on the imputed genotyped dosages (maf >0.01, HWE-p.value >1 × 10^{-6}, imputation R^2 >0.3) for bi-allelic autosomal markers. We performed all the GWAS runs under the linear mixed model framework implemented in BOLT-LMM [9].

We applied the SNP2GENE function from FUMA [10] on the summary statistics to identify genomic risk loci for the phenotype. For each locus, we selected the top lead SNP (single nucleotide polymorphism) from a set of independent genome-wide significant SNPs ($p \leq 5\times10^{-8}$) within a 250kb LD-block window (r^2 <0.5). These SNPs were then used for downstream analysis.

We repeated the analysis in participants who identified as Black/Black British, and then in participants who identified as Asian/Asian British & Chinese, separately. We compared the effect

sizes of the genome-wide significant associations in White British with those of Black/Black British and Asian/Asian British & Chinese (Figure S1). We find, as expected [11, 12, 13], that results are broadly correlated, but that the effect sizes are systematically lower in Black/Black British populations, sometimes to the point that differences are significant for individual variants (i.e., effect sizes are inconsistent). Such an analysis is straightforward once a GWAS pipeline has been set up and is a first line of evidence about the replicability of associations.



(a) (b)

Figure S1: **Comparison of effect sizes across populations for phenotype 'eosinophil cell counts' and genetic variants reaching significance in the majority population.**

(a) White British and Asian/Asian British & Chinese populations, slope C.I. [0.63 - 1.06], r = 0.57 (b) White British and Black/Black British populations; slope C.I. [0.45 - 0.73], r = 0.48. Each dot represents a significant variant from a White British GWAS (top lead SNP) and for which standard error in *MIN* is below 0.02. 23.4% of the variants replicated at a 0.05 p-value in the Asian/Asian British & Chinese population compared to 7.1% in the Black/Black British population (blue dots). The red dot is a variant with inconsistent effect (p < 0.05 in *MIN* but with opposite effect direction).

5. Software availability

Code used to generate the results in this manuscript can be accessed here: https://github.com/Chiefeghan/GWAS.

6. Data availability

UK Biobank analyses were conducted via application 6728.

UK Biobank summary statistics from our analyses are publicly available at : https://github.com/Chiefeghan/GWAS.

7. References

[1] Anna Fry, Thomas J. Littlejohns, et al. “Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population”. *American Journal of Epidemiology* 186.9 (2017), 1026.

[2] Annalisa Buniello, Jacqueline A L MacArthur, et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. *Nucleic Acids Research* 47.D1 (2018, accessed February 2019), p. D1005.

[3] Milly S. Tedja, Robert Wojciechowski, et al. “Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error”. *Nature Genetics* 50.6 (2018), p. 834.

[4] Helen R. Warren, Evangelos Evangelou, et al. “Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk”. *Nature Genetics* 49.3 (2017), p. 403.

[5] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. “Rayyana web and mobile app for systematic reviews”. *Systematic Reviews* 5.1 (2016), p. 210.

[6] Health and Retirement Study Survey Publications. (accessed February 2019, https://hrs.isr.umich.edu/publications/biblio/).

[7] William J. Astle, Heather Elding, et al. “The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease”. *Cell* 167.5 (2016), p. 1415.

[8] Jan Ltvall, Cezmi A. Akdis, et al. “Asthma endotypes: Anew approach to classification of disease entities within the asthma syndrome”. *Journal of Allergy and Clinical Immunology* 127.2 (2011), p. 355.

[9] Po-Ru Loh, George Tucker, et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. *Nature Genetics* 47.3 (2015), p. 284.

[10] Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle Posthuma. “Functional mapping and annotation of genetic associations with FUMA”. *Nature communications* 8.1 (2017), p. 1826.

[11] Urko M. Marigorta and Arcadi Navarro. “High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants”. *PLOS Genetics* 9.6 (2013), e1003566.

[12] Genevieve L. Wojcik, Mariaelisa Graff, et al. “Genetic analyses of diverse populations improves discovery for complex traits”. *Nature* 570.7762 (2019), p. 514.

[13] Rachel A Myers, Blanca E Himes, et al. “Further replication studies of the EVE Consortium meta-analysis identifies 2 asthma risk loci in European Americans”. *Journal of Allergy and Clinical Immunology* 130.6 (2012), p. 1294.